DOCUMENT RESUME

ED 430 182 CG 029 269

AUTHOR Hummel, Thomas J.; Lichtenberg, James W.

TITLE Limits of Meta-Analysis as a Basis for Justifying Individual Counseling Interventions.

PUB DATE 1999-04-23

NOTE 20p.; Paper presented at the Annual Conference of the American Educational Research Association (Montreal, Quebec, Canada, April 19-23, 1999).

PUB TYPE Reports - Evaluative (142) -- Speeches/Meeting Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.

DESCRIPTORS *Counseling; *Intervention; *Meta Analysis; Models; Outcomes of Treatment; Psychotherapy; Research

ABSTRACT

The first part of this paper examines using meta-analysis as a basis for making probabilistic statements about client outcome by revisiting Smith and Glass's (1977) classic paper concerning meta-analysis of psychotherapy outcome studies. While it may be argued that making probabilistic statements was not the primary purpose of meta-analysis, Smith and Glass took it in that direction when they presented a figure depicting two overlapping normal distributions, one representing the treated population and the other representing the control population. They pointed out that a person at the mean of the treated population fell at the 75th percentile of the control group. From this it can be deduced that the probability is 0.75 of an individual randomly drawn from the treated population being above the mean of the control population. This paper assesses the justification for this statement using numerical analysis and model fitting techniques. Results show that normal distribution is probably not an appropriate model for treated subjects and working backwards from effect size distributions to client distributions seems doomed to failure. The second part of the paper considers the kinds of probabilistic statements that might be offered to clients and how three models of client outcomes might relate to the statements. Six figures depict the different model analyses. (MKA)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 430 182

Limits of meta-analysis as a basis for justifying individual counseling interventions

Thomas J. Hummel
University of Minnesota

James W. Lichtenberg
University of Kansas

Paper presented at the annual meeting of the American Educational Research Association, April 23, 1999, Montreal, CANADA.

BEST COPY AVAILABLE

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

This document has been reproduced as received from the person or organization originating it.

Minor changes have been made to improve reproduction quality.

Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

J. LICHTENBERG

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

CG029269

Limits of meta-analysis as a basis for justifying individual counseling interventions

This paper rests on the assumption that a counseling client is a person who is motivated to change and willing to invest time and money pursuing counseling if it will achieve a desired outcome. Clients are seen as consumers who have a right to know the benefits they will receive. We believe that many clients want an answer to the following question: "Am I going to get better?" We see the answer to this question as being probabilistic in nature, communicating the likelihood of improvement, or the likelihoods of various degrees of improvement. From our perspective, mean differences between counseled and uncounseled groups and the effect sizes derived from them are not an adequate basis for a client's decision to begin counseling.

A reasonable indicator of how a particular client will do in counseling could be based on the outcomes of similar clients. If counseling outcome varies as a function of client characteristics, then we have to have sufficient outcome data on each type of client to make a reliable statement about client improvement. Individual studies do not have enough data to make reliable statements about individual client outcome. Therefore, we must aggregate data across studies. Currently, the most widely used method of aggregation is meta-analysis.

The first part of the paper looks at using meta-analysis as basis for making probabilistic statements about client outcome by revisiting Smith and Glass's (1977) classic paper. While it may be argued that making probabilistic statements was not the primary purpose of meta-analysis, Smith and Glass took it in that direction when they presented a figure depicting two overlapping normal distributions, one representing the population of those who were treated and the other representing those in the control population. They pointed out that a person at the mean of the treated population fell at the 75th percentile of the control group. From this it can be deduced that the probability is 0.75 of an individual randomly drawn from the treated population being above the mean of the control population.[1] We will assess the justification for this statement using numerical analysis and model fitting techniques.

Are probabilistic outcome statements based on Smith and Glass's 1977 findings justified?

[1] Given the symmetry of the normal distribution, if the mean of those treated falls at the 75th percentile of the control population, then the mean of the control population must fall at the 25th percentile of the treated population. Therefore, 75% of the those treated must fall above the mean of the control population.

The answer to the preceding question rests in large part on how well the normal distributions presented model Smith and Glass's (1977) data. While admitting that there was no justification for using normal distributions, Smith and Glass stated that "normality has as much justification as any other form" (p. 754). We disagree with their contention, primarily because the effect size distribution had a skewness of 0.99.

It is not surprising that the effect size distribution is skewed to some extent. If the underlying populations of clients were normally distributed, then effect sizes would follow the non-central *t*-distribution, which (in this case) is positively skewed. The question is: "Is the skewness of the non-central *t*-distribution sufficient to explain the skewness in Smith and Glass's distribution of effect sizes?" In our attempt to answer this question, and others posed in this section, the distribution theory we used assumed independently distributed random variables. While the results from the 375 different studies included in the Smith and Glass analysis could reasonably be assumed to be independent, multiple effect sizes from the same study could not be. The effect of this partial dependence is unknown.

The non-central *t*-distribution as an explanation for the degree of skewness in the effect size distribution: The skewness of the non-central *t*-distribution increases as the non-centrality parameter increases and as the degrees of freedom decrease. While there are non-central *t*-distributions with a skewness of 0.99, given the effect sizes and the sample sizes typically reported in the studies Smith and Glass used, the skewness of the non-central *t*-distribution would be far less than that of the effect size distribution. For example, with an effect size of ES = 0.68, which is equal to the overall effect size reported by Smith and Glass, and degrees of freedom of 20, which we believe to be at the low end for their studies, the skewness of the non-central *t*-distribution would be 0.34. Further, the proportion of negative effect sizes and the standard deviation of effect sizes do not agree with what would be expected if the effect size distribution were a linear transformation of the non-central *t*-distribution. For these reasons, the non-central *t*-distribution is not a good explanatory model for the distribution of effect sizes or its skewness.

If the non-central *t*-distribution is not a good model for the effect size distribution, perhaps some function of it is. We speculated that highly positive research findings, i.e., large positive effect sizes, would be more likely to see the light of day as a publication or presentation than highly negative findings. We postulated a type of censorship, perhaps imposed by the original researcher or someone else, that increased the likelihood that a study would be included in a meta-analysis as the effect size for the study became more supportive of treatment effectiveness. If this model proved adequate, then one could remove the censorship and see what the effect size distribution would have looked like had all studies been included in the meta-analysis. With the average effect size from this "reconstructed" distribution, the overlapping normal distributions could be redrawn and the probabilistic statements of effectiveness revised.

Censorship as an explanation for the degree of skewness in the effect size distribution: The model we used was probabilistic and assumed that the larger the effect size the more

likely it would get into a meta-analysis. This had the effect of decreasing the frequency of values in the left-hand tail of the effect size distribution, thereby increasing positive skewness. This approach has to be taken cautiously because Smith and Glass went to great lengths to ensure that all findings, published or not, were included. The probability models that led to distributions approximating Smith and Glass's distribution of effect sizes had about 50% of the "original" effect sizes missing. That half the studies would be missing seems extremely unlikely given the very thorough procedures that Smith, Glass and Miller (1980) reported for acquiring studies. For this reason, censorship did not provide an adequate explanation.

We next entertained the proposition that the overlapping normal distribution model was appropriate for sub-populations of clients, while not fitting the overall population. After all, Smith and Glass reported average effect sizes that varied from 0.26 to 0.91 for different approaches to therapy. Perhaps mixing the various therapies together produced the skewness in the overall effect size distribution.

Mixtures of non-central *t*-distributions as an explanation for the degree of skewness in the effect size distribution: We let type of therapy define sub-populations of clients. Smith and Glass reported the average effect size, the number of effect sizes, and the standard error of the mean effect size for each of ten therapies. With this information, we formed a mixture of ten non-central *t*-distributions, letting each distribution be weighted according to its proportion of the total number of effect sizes.[2] The resulting distribution's mean, standard deviation, and proportion of negative effect sizes approximated to a reasonable degree the corresponding values reported by Smith and Glass. The mixture, however, had a skewness of 0.45, which is far less than the.0.99 reported by Smith and Glass.

The last explanation we considered is that the underlying distributions of treated and control subjects do not have the same shape. Since the previous three sources of skewness seem insufficient as an explanation of the skewness in the effect size distribution, we were led to consider that at least part of the skewness came from the distribution of the original data. This, of course, implies that the overlapping normal distribution model is incorrect.

---

[2] We noted that the number of effect sizes reported for the ten therapies totaled 744, while the article reported 833 effect sizes. It seemed likely that the "missing" effect sizes belonged to a "placebo" category. This implied that there was a missing category containing 89 effect sizes. Since we knew the mean and standard deviation of the 833 effect sizes, and we could determine the mean and standard deviation of ten of the 11 categories, we attempted to solve for the 11th group's mean and variance. We solved first for the mean, and then, using this value, we solved for the variance. The solution was negative. Since variances are positive, it was clear that there was an error. After carefully checking our computations, we believe that the error is in the reported results, e.g., maybe there is a typographical error among the reported mean effect sizes for the therapies. Since we could not obtain values for the 11th group, we decided to proceed with the ten therapies for which results were reported. Since our results are based on findings that we believe contain an error, we present them as tentative. On the positive side, none of Smith and Glass' results are "out of the ballpark," and we assume that the correct values would not substantially change our conclusions.

The answer to the preceding question rests in large part on how well the normal distributions presented model Smith and Glass's (1977) data. While admitting that there was no justification for using normal distributions, Smith and Glass stated that "normality has as much justification as any other form" (p. 754). We disagree with their contention, primarily because the effect size distribution had a skewness of 0.99.

It is not surprising that the effect size distribution is skewed to some extent. If the underlying populations of clients were normally distributed, then effect sizes would follow the non-central *t*-distribution, which (in this case) is positively skewed. The question is: "Is the skewness of the non-central *t*-distribution sufficient to explain the skewness in Smith and Glass's distribution of effect sizes?" In our attempt to answer this question, and others posed in this section, the distribution theory we used assumed independently distributed random variables. While the results from the 375 different studies included in the Smith and Glass analysis could reasonably be assumed to be independent, multiple effect sizes from the same study could not be. The effect of this partial dependence is unknown.

The non-central *t*-distribution as an explanation for the degree of skewness in the effect size distribution: The skewness of the non-central *t*-distribution increases as the non-centrality parameter increases and as the degrees of freedom decrease. While there are non-central *t*-distributions with a skewness of 0.99, given the effect sizes and the sample sizes typically reported in the studies Smith and Glass used, the skewness of the non-central *t*-distribution would be far less than that of the effect size distribution. For example, with an effect size of ES = 0.68, which is equal to the overall effect size reported by Smith and Glass, and degrees of freedom of 20, which we believe to be at the low end for their studies, the skewness of the non-central *t*-distribution would be 0.34. Further, the proportion of negative effect sizes and the standard deviation of effect sizes do not agree with what would be expected if the effect size distribution were a linear transformation of the non-central *t*-distribution. For these reasons, the non-central *t*-distribution is not a good explanatory model for the distribution of effect sizes or its skewness.

If the non-central *t*-distribution is not a good model for the effect size distribution, perhaps some function of it is. We speculated that highly positive research findings, i.e., large positive effect sizes, would be more likely to see the light of day as a publication or presentation than highly negative findings. We postulated a type of censorship, perhaps imposed by the original researcher or someone else, that increased the likelihood that a study would be included in a meta-analysis as the effect size for the study became more supportive of treatment effectiveness. If this model proved adequate, then one could remove the censorship and see what the effect size distribution would have looked like had all studies been included in the meta-analysis. With the average effect size from this "reconstructed" distribution, the overlapping normal distributions could be redrawn and the probabilistic statements of effectiveness revised.

Censorship as an explanation for the degree of skewness in the effect size distribution: The model we used was probabilistic and assumed that the larger the effect size the more

likely it would get into a meta-analysis. This had the effect of decreasing the frequency of values in the left-hand tail of the effect size distribution, thereby increasing positive skewness. This approach has to be taken cautiously because Smith and Glass went to great lengths to ensure that all findings, published or not, were included. The probability models that led to distributions approximating Smith and Glass's distribution of effect sizes had about 50% of the "original" effect sizes missing. That half the studies would be missing seems extremely unlikely given the very thorough procedures that Smith, Glass and Miller (1980) reported for acquiring studies. For this reason, censorship did not provide an adequate explanation.

We next entertained the proposition that the overlapping normal distribution model was appropriate for sub-populations of clients, while not fitting the overall population. After all, Smith and Glass reported average effect sizes that varied from 0.26 to 0.91 for different approaches to therapy. Perhaps mixing the various therapies together produced the skewness in the overall effect size distribution.

Mixtures of non-central *t*-distributions as an explanation for the degree of skewness in the effect size distribution: We let type of therapy define sub-populations of clients. Smith and Glass reported the average effect size, the number of effect sizes, and the standard error of the mean effect size for each of ten therapies. With this information, we formed a mixture of ten non-central *t*-distributions, letting each distribution be weighted according to its proportion of the total number of effect sizes.[2] The resulting distribution's mean, standard deviation, and proportion of negative effect sizes approximated to a reasonable degree the corresponding values reported by Smith and Glass. The mixture, however, had a skewness of 0.45, which is far less than the.0.99 reported by Smith and Glass.

The last explanation we considered is that the underlying distributions of treated and control subjects do not have the same shape. Since the previous three sources of skewness seem insufficient as an explanation of the skewness in the effect size distribution, we were led to consider that at least part of the skewness came from the distribution of the original data. This, of course, implies that the overlapping normal distribution model is incorrect.

---

[2] We noted that the number of effect sizes reported for the ten therapies totaled 744, while the article reported 833 effect sizes. It seemed likely that the "missing" effect sizes belonged to a "placebo" category. This implied that there was a missing category containing 89 effect sizes. Since we knew the mean and standard deviation of the 833 effect sizes, and we could determine the mean and standard deviation of ten of the 11 categories, we attempted to solve for the 11th group's mean and variance. We solved first for the mean, and then, using this value, we solved for the variance. The solution was negative. Since variances are positive, it was clear that there was an error. After carefully checking our computations, we believe that the error is in the reported results, e.g., maybe there is a typographical error among the reported mean effect sizes for the therapies. Since we could not obtain values for the 11th group, we decided to proceed with the ten therapies for which results were reported. Since our results are based on findings that we believe contain an error, we present them as tentative. On the positive side, none of Smith and Glass' results are "out of the ballpark," and we assume that the correct values would not substantially change our conclusions.

Different distributions for treated and control subjects as an explanation for the degree of skewness in the effect size distribution: It can be shown that if the underlying distributions of treated and control subjects were both symmetrical, then the sampling distribution of mean differences between treated and control groups would also be symmetrical. Further, it can be shown that if the underlying distributions of treated and control subjects were both skewed in precisely the same manner, and if they had the same variance and samples drawn from them were the same size, then the sampling distribution of mean differences between treated and control groups would still be symmetrical. To simplify matters, we assumed that samples drawn from the treated and control population were the same size. Granting this simplification, the underlying distributions will only cause the sampling distribution of mean differences to be skewed if 1) at least one of the distributions is skewed and 2) the distributions are not identical in skewness and variance. As a way to meet these conditions, we continued to assume that the control population was normally distributed and assumed a different distribution for the treated population. To accomplish this, we assumed that treatment effectiveness varied across clients, i.e., that therapy was more beneficial for some than it was for others. Second, we assumed that while therapy can sometimes be of great benefit, it is relatively less likely to do great harm. For example, suppose therapy outcome were rated on a seven-point scale from "extremely negative" to "extremely positive." Our assumption was that the frequency of "extremely negative" would be less than the frequency of "extremely positive." This last assumption leads to a positively skewed distribution of individual effect sizes for the treated population. The control subjects' values simply represented the "natural" variability among untreated subjects, and, as stated, these values were assumed to be normally distributed. The treated subjects' values were assumed to be constructed of two, independently distributed additive parts, one identical in distribution to that of the control subjects and a second part randomly sampled from a skewed distribution of individual treatment effects. Varying the skewness and variance of the distribution of individual treatment effects affected the skewness and variance of the treated population, which, in turn, affected the skewness and variance of the sampling distribution of the difference between the control and treatment means. Given the assumptions we have made, the functional relationship between the skewness of the treated population and the skewness of the sampling distribution of the difference between the control and treatment means is such that the skewness in the mean difference distribution will be much less than the skewness in the treated population. Even when we increased the variance and skewness of the distribution of individual treatment effects to what we considered an upper limit, the skewness of the distribution of mean differences rose to only 0.12. At this point, we do not know what the effect size distribution would be when the distribution of mean differences is skewed to this extent (0.12). The underlying distributions that lead to this skewness violate the assumptions of the non-central *t*-distribution, so that distribution is no longer appropriate. Our intuition is that this rather modest degree of skewness, namely, 0.12, would be very unlikely to lead to the effect size skewness of 0.99 observed by Smith and Glass.

We have concluded from the preceding four subsections that none of the sources of skewness considered begins to explain the skewness in Smith and Glass' data. Perhaps if combined in some manner, the sources could come closer to 0.99. Given the explanatory

power of each source, though, we are somewhat skeptical that any combination would be adequate.

To us, the preceding analysis demonstrates two things: first, the normal distribution is probably not an appropriate model for treated subjects; and second, working backwards from effect size distributions to client distributions seems doomed to failure. If we want to make probabilistic statements about client outcomes, we will have to take a different approach than meta-analysis.

Next, we will consider the kinds of probabilistic statements we might wish to make to clients and how various models of client outcome might relate to these statements. We will consider meta-analysis in the following discussion, for it provides a familiar starting point; but in addition to it, we will discuss other approaches.

Probabilistic client outcomes and models of counseling effectiveness

In the beginning of this paper we suggested that clients might want an answer to the question, "Am I going to get better?", and our focus on probabilistic statements indicates that the response to this question would not be a simple "Yes" or No." An answer must relate to a way to compute a probability. To help in conceptualizing ways to compute probabilities, we will consider three related, but more formal, questions. These questions assume an outcome measure on which higher values are better than lower values:

1. What is the probability that a person randomly chosen from those that have been counseled will score higher than the mean of the population of those not counseled?

2. What is the probability that a person randomly chosen from the counseled population will be higher than a person randomly drawn from the population of those not counseled?

3. What is the probability that a randomly chosen person from the counseled population will improve?"

All three questions would be answered with a probability of some outcome, and all of the outcomes relate in some way to the effectiveness of counseling. The first question derives from the standard meta-analysis model, and the probability changes as a function of the effect size (ES). We will denote this probability based on the effect size as PES. The second question relates to what has been called the "probability of superiority," (PS) the probability that a treated person will have a better outcome than a non-treated person (Grissom, 1996). The last probability deals with what we will call the probability of improvement, or PI. This is simply the probability that a person will change for the better. In the following, we will compare these three probabilities, PES, PS, and PI, using different models of treatment effect.

We will first consider three simple models of treatment effectiveness. The first model is consistent with traditional meta-analysis and assumes that that each treated client receives an identical benefit. The second model introduces individual effect sizes, i.e., clients can have different reactions to treatment. The third model suggests a simple relationship between individual differences and individual effect sizes. For brevity's sake, we will often refer to those who have been counseled or have received therapy as "treated" and those who have not as "control." We will use the normal distribution in our presentation, not because we think it is an adequate model (see above), but because it is simple and useful for comparing various probabilities. At the end of this section, we will discard the normal model and discuss more realistic approaches for estimating probabilities.

Model I: Control: $Y_c = \varepsilon \sim N(0,1)$; Treated: $Y_t = \delta + \varepsilon \sim N(\delta,1)$, where $\delta \geq 0$[3].

In Model I, the control population is normally distributed with a mean of zero and a standard deviation of one, while the treated population has the same distribution except for the mean, δ. We might think of ε as representing individual differences on some counseling relevant dimension, such as depression. Throughout the following we will think of low (negative) values of ε as being more problematic, for example, as representing increased depression, so that higher scores are better. This convention leads to positive values for effect sizes when the treatment is helpful, which is usually the case when reporting meta-analyses. Given that the control population is represented by the standard normal distribution, δ is equal to the population value of the effect size associated with treatment, e.g., the benefit due to cognitive therapy for depression. If $\delta = 0.7$, then the following picture represents control and treatment populations.

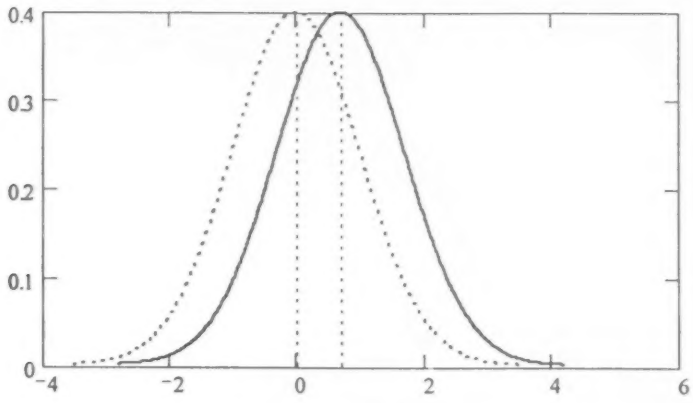

Figure 1. Meta-Analysis

[3] The general notation $N(\mu,\sigma^2)$ refers to a normal distribution with a mean of μ and a variance of σ^2.

Figure 1 represents the traditional meta-analysis picture. While Model I need not be assumed when computing effect sizes, it describes the pictures of overlapping distributions that accompany effect size presentations. Further, the model can be used to make statements like, "the average person who receives therapy is better off at the end of it than 80 percent of the persons who do not" (Smith, Glass, & Miller, 1980, p. 87). The model postulates that the treatment and control differ only by a constant that shifts the treatment distribution. This simple model is consistent with the linear model taught in analysis of variance texts.

In Figure 1, with $\delta = 0.7$, the mean of the treated population falls at the 76th percentile of the control population. Viewed another way, we could say that the 24th percentile of the treatment group falls at the mean of the control, and since the control group's mean is zero, this view leads to the following probability statement: $Pr[Y_t > 0] = 0.76$. This probability says that if you are treated, the probability is 0.76 that you will do better than the mean of the control group. This probability that a person randomly chosen from treated population will score higher than the mean of the non-treated population is an answer to question one above. It is the probability that we have denoted as PES.

Using Model I, we can also define the probability of superiority, PS. PS is the probability that a person randomly drawn from the treated population will have a higher dependent variable value than a randomly drawn person from the control population, or $Pr[Y_t > Y_c] = Pr[Y_t - Y_c > 0]$. For $\delta = 0.7$, PS = 0.69.

Y_t is distributed $N(\delta,1)$ and $Y_t - Y_c$ is distributed as $N(\delta,2)$. If we transform Y_t and $Y_t - Y_c$ to the standard normal distribution[4], then the following two integrals define the probabilities PES and PS:

$$PES = \frac{1}{\sqrt{2\pi}} \int_{-\delta}^{+\infty} e^{-\frac{1}{2}z^2} dz$$

$$PS = \frac{1}{\sqrt{2\pi}} \int_{\frac{-\delta}{\sqrt{2}}}^{+\infty} e^{-\frac{1}{2}z^2} dz$$

From the above integrals it is clear that if $\delta > 0$, PES is greater than PS because the following integral, which evaluates the difference between PES and PS, is always greater than zero:

$$PES - PS = \frac{1}{\sqrt{2\pi}} \int_{-\delta}^{\frac{-\delta}{\sqrt{2}}} e^{-\frac{1}{2}z^2} dz$$

[4] Using the standard normal distribution allows us to compare PES and PS by defining a region of that distribution that we can integrate to find the difference between PES and PS.

PES and PS are graphed as a function of δ in the following figure:

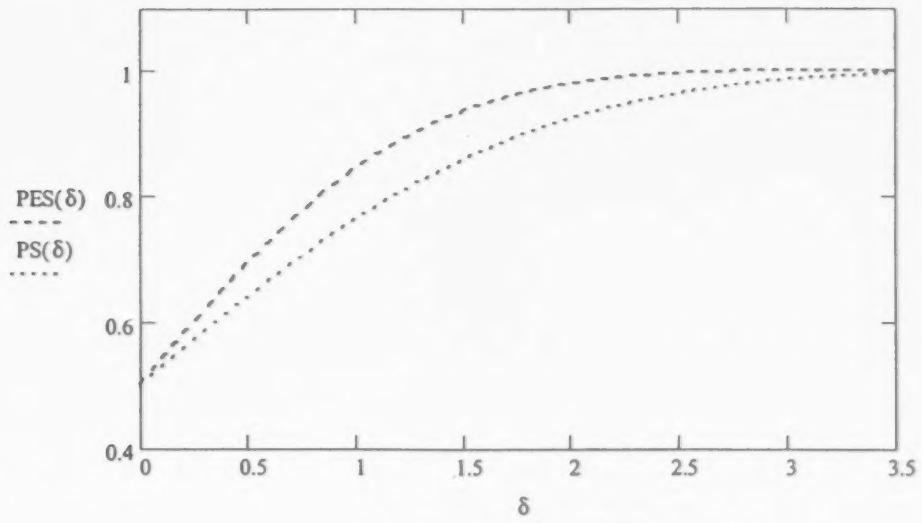

Figure 2. PES and PS

The maximum difference between PES and PS is PES – PS = 0.083 and it occurs at an effect size of δ = 1.18. Effect sizes in the interval 0.47 to 2.11 all give differences in excess of 0.05.

For Model I, the probability of improvement, PI, is equal to 1.00, because everyone in the treated population improved by δ. For a client who wants to know the probability that they will get better, PES and PS substantially underestimate PI for all except the largest values of δ.

We think that a client's conversational versions of the above three questions we posed might be "What are the chances I'll do better than the average untreated person?" (PES), "What are the chances I'll be better off than someone who's not counseled?" (PS), and "What are the chances I'll be better off than if I don't get counseling?" (PI). In all three questions, the client is wondering if he or she will do better and a counselor's response could be a probability, which, at this point, would only take into account the fact that a person would be counseled. What changes across the questions is the reference point, to what is the client being compared, i.e., the client will do better than what? In our opinion, the last question seems the most germane to clients because the reference point is the client, not the mean of a group or some unknown person. "What are the chances I'll be better off than if I don't get counseling?" means "What are the chances I'll improve?". If one agrees with our opinion that the last question is the most relevant of the three, then PES and PS have value only in so far as they approximate PI, and, for the most part, PES and PS fail in this respect.

Solely from a marketing perspective, one would like to give the client the highest probability of improvement possible, for all else being equal, the higher the probability of improvement, the more likely the client would be to choose counseling. While we are certainly not suggesting that marketing issues be considered, we believe that counselors who agree with us that PI is the most germane to a client's decision should not shy away from PI simply because it leads to a marketing advantage.

In our discussion of Model I, PI = 1.00, because all clients improve. Further, they all improve by the same amount. However, in our discussions with counselors and counselor trainees, we have found no one who believes this to be a realistic model of counseling outcome. They believe that clients vary in their response to counseling and that some clients might deteriorate in counseling rather than improve. The next model, Model II, extends Model I by allowing individual counseling outcomes.

Model II: Control: $Y_c = \varepsilon \sim N(0,1)$; Treated: $Y_t = \delta + \varepsilon \sim N(\mu_\delta, 1+\sigma_\delta^2)$, where $\mu_\delta \geq 0$, $\delta \sim N(\mu_\delta, \sigma_\delta^2)$, and $\rho_{\delta,\varepsilon} = 0$.

In Model II, the control population remains normally distributed with a mean of zero and a standard deviation of one, while the treated population's distribution results from adding two independent random variables, δ, an individual's treatment effect, and ε, the sampling error. If $\mu_\delta = 0.7$ and $\sigma_\delta = 0.5$, then the following picture represents this situation.

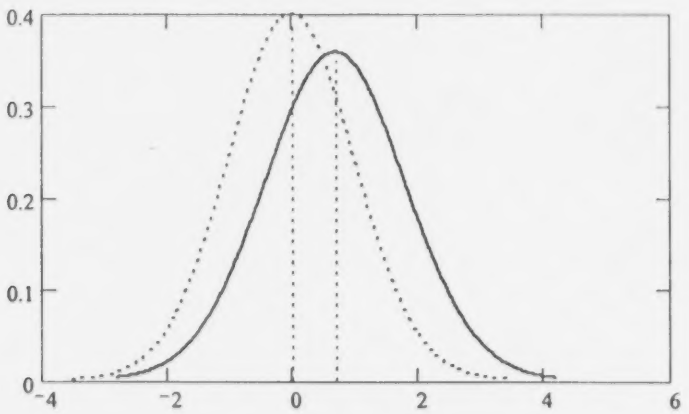

Figure 3. Variable Effect Sizes

The treated population is now more dispersed than the control because its standard deviation has increased to $\sigma_t = 1.12$ due to the variability of the individual effect sizes. This change from Figure 1 above affects PS, but may not affect PES. If a researcher used the original definition of effect size, using the standard deviation of the control group in the denominator, then the increased variability of the treated population would not affect the effect sizes. For simplicity, we will assume the effect size is based on the original

definition.[5] The reason PES remains unchanged is because the treatment and control populations are depicted as differing only as a function of the average effect size, which in this case is equal to μ_δ.

Even though Model II causes Figure 3 to differ from Figure 1, the distributions in Figure 1 would be used to compute PES. The difference between the treated populations depicted in Figure 1 and Figure 3 will therefore cause PES to be in error, because Y_t no longer has a standard deviation of 1.00. We will not consider the correct probability for PES = Pr[$Y_t > 0$] because that is not the probability that would be reported given the usual meta-analysis depiction.

The probability of superiority, PS = $Pr[Y_t > Y_c] = Pr[Y_t - Y_c > 0]$, changes because Y_t is now distributed $N(\mu_\delta, 1 + \sigma_\delta^2)$ and $Y_t - Y_c$ is distributed as $N(\mu_\delta, 2 + \sigma_\delta^2)$. If we transform $Y_t - Y_c$ to the standard normal distribution, then the following integral defines PS:

$$PS = \frac{1}{\sqrt{2\pi}} \int_{\frac{-\mu_\delta}{\sqrt{2+\sigma_\delta^2}}}^{+\infty} e^{-\frac{1}{2}z^2} dz.$$

In Model I above, the probability of improvement, PI, equaled 1.00, because everyone in the treated population improved by the same amount, δ. Here, however, δ is a random variable distributed as $N(\mu_\delta, \sigma_\delta^2)$. If we transform δ to the standard normal distribution, then the following integral defines PI = $Pr[\delta > 0]$:

$$PI = \frac{1}{\sqrt{2\pi}} \int_{\frac{-\mu_\delta}{\sigma_\delta}}^{+\infty} e^{-\frac{1}{2}z^2} dz.$$

For $\mu_\delta = 0.7$ and $\sigma_\delta = 0.5$, the values on which Figure 3 is based, PS = 0.68 and PI = 0.92. Again, the interpretation of PI seems more germane than PS, for PI tells the client that the probability is 0.92 of an improvement, and 0.08 that the client will get worse. For purposes of comparison, PES = 0.76. This value is found using the Model I.

One can determine that if $\mu_\delta > 0$, PI is always greater than PS, because the following integral, which evaluates the difference between PI and PS, is always greater than zero:

$$PI - PS = \frac{1}{\sqrt{2\pi}} \int_{\frac{-\mu_\delta}{\sigma_\delta}}^{\frac{-\mu_\delta}{\sqrt{2+\sigma_\delta^2}}} e^{-\frac{1}{2}z^2} dz.$$

[5] This assumption is rather inconsequential for the parameters with which we are working. If the "pooled" variance where used to obtain the standard deviation, then the standard deviation would be 1.061 and the population effect size would drop from 0.70 to 0.66. This change would drop PES from 0.76 to 0.75.

The following figure compares PS and PI, for $\sigma_\delta = 0.5$ and $\mu_\delta = 0.0 \ldots 3.50$. PES based on the Model I definition of ES is also included.

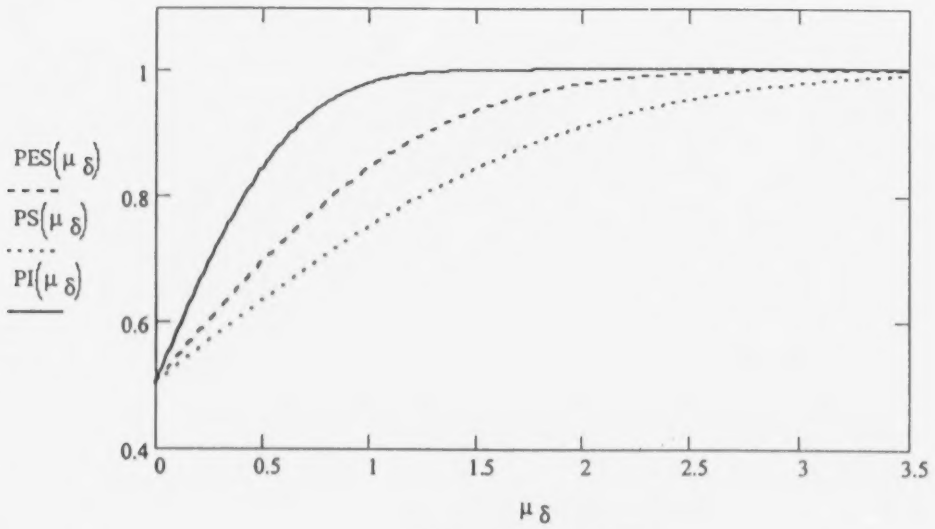

Figure 4. PES, PS, PI

For the values plotted above, a maximum difference of 0.16 between PI and PES occurs at $\mu_\delta = 0.68$. At this point, the difference between PI and PS is 0.24. For what we judge to be likely values of average individual effect size, say, 0.5 to 1.5, we consider PES and PS to differ in an important manner from PI.

Model II assumes that δ and ε are independent. For example, if ε represents a client's level of self esteem without counseling, then whether clients are high or low on self esteem bears no relationship to the amount of benefit they would receive from counseling. Clearly, other models are possible. Perhaps, a client higher on self esteem cannot expect to make gains as big as a person lower on self esteem. The next model includes such a relationship.

Model III: **Control:** $Y_c = \varepsilon \sim N(0,1)$; **Treatment:** $Y_t = \delta + \varepsilon \sim N(\mu_\delta, 1+\sigma_\delta^2+2\rho_{\delta,\varepsilon}\sigma_\delta)$, **where** $\mu_\delta \geq 0$, $\delta \sim N(\mu_\delta, \sigma_\delta^2)$, **and** $\rho_{\delta,\varepsilon} < 0$.

In Model III, the control population remains normally distributed with a mean of zero and a standard deviation of one, while the treated population's distribution results from adding two correlated random variables, δ, an individual treatment effect, and ε, the sampling error. If $\mu_\delta = 0.7$, $\sigma_\delta = 0.5$, and $\rho_{\delta,\varepsilon} = -0.5$, then the following picture represents this situation.

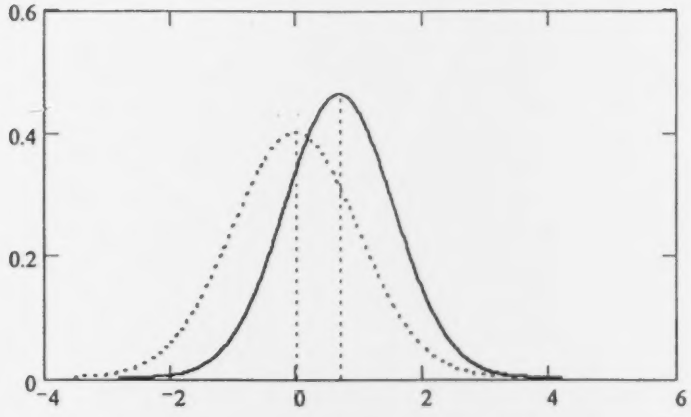

Figure 5. ESs Correlated with Error

In contrast to Model I and Model II, Figure 5 depicts less variability in the treated population. This is due to the negative correlation between δ and ε, which causes σ_t to decrease to 0.87.

For reasons presented above in discussing Model I, PES is unaffected because the changes introduced in Model III do not affect the control population. PI is not affected by the model changes either, because the distribution of δ is unaffected. PS is affected, however, by the reduction in σ_t.

The graph of PES, PS, and PI for Model III, with $\sigma_\delta = 0.5$ and $\rho_{\delta,\varepsilon} = -0.5$, would be very similar to Figure 4 for Model II. As just stated, PES and PI are unaffected by the introduction of $\rho_{\delta,\varepsilon} = -0.5$ and their curves would be unchanged. PS's curve would stay in the same relative position in relation to PES and PI, but for most values of μ_δ it would be slightly elevated. The largest difference between PS under Model II and Model III would occur at around $\mu_\delta = 1.4$. The difference there would be 0.03, with the differences decreasing on either side of 1.4. The discussion of PS under Model III, therefore, would be substantively the same as under Model II.

With the introduction of $\rho_{\delta,\varepsilon} < 0$, Model III allows for an entirely new line of analysis based on conditional probability. Instead of thinking about a randomly drawn client from the treated population, we can condition on information about the individual, in this case, ε. Clients could be given probabilities based on clients that are similar to them in a given manner.

We will pursue conditional probabilities only for PI, the probability of improvement. Published methods that we have reviewed for meta-analysis and computing the probability of superiority, PS, do not deal with estimating conditional probabilities. These approaches could take individual differences into account by blocking studies (or effect sizes or subjects) into subsets defined by particular subject attributes, e.g., a subset for upper class, African-American females with four-year college degrees. However, if this were done, the number of elements in the subsets would decrease, and estimates of treatment effect based on the subset would become less reliable. We will leave it to others more committed to these approaches (PES and PS) to pursue the topic of conditional probabilities.

For the probability of improvement, we will define PI' $= \Pr[\delta > 0 \mid \varepsilon]$. This conditional probability of improving depends on one's standing on the dependent variable prior to the addition of the treatment effect. Since we are using $\rho_{\delta,\varepsilon} = -0.5$, the model predicts that clients with smaller values of ε (i.e., clients with more serious problems) have a higher likelihood of improving during counseling. For $\mu_\delta = 0.7$, $\sigma_\delta = 0.5$, and $\rho_{\delta,\varepsilon} = -0.5$, the following figure displays curves for PI' for the following values of ε: -2, -1, 0, 1, 2. A curve for PES based on Model I and one for PS based on Model III are included for reference.

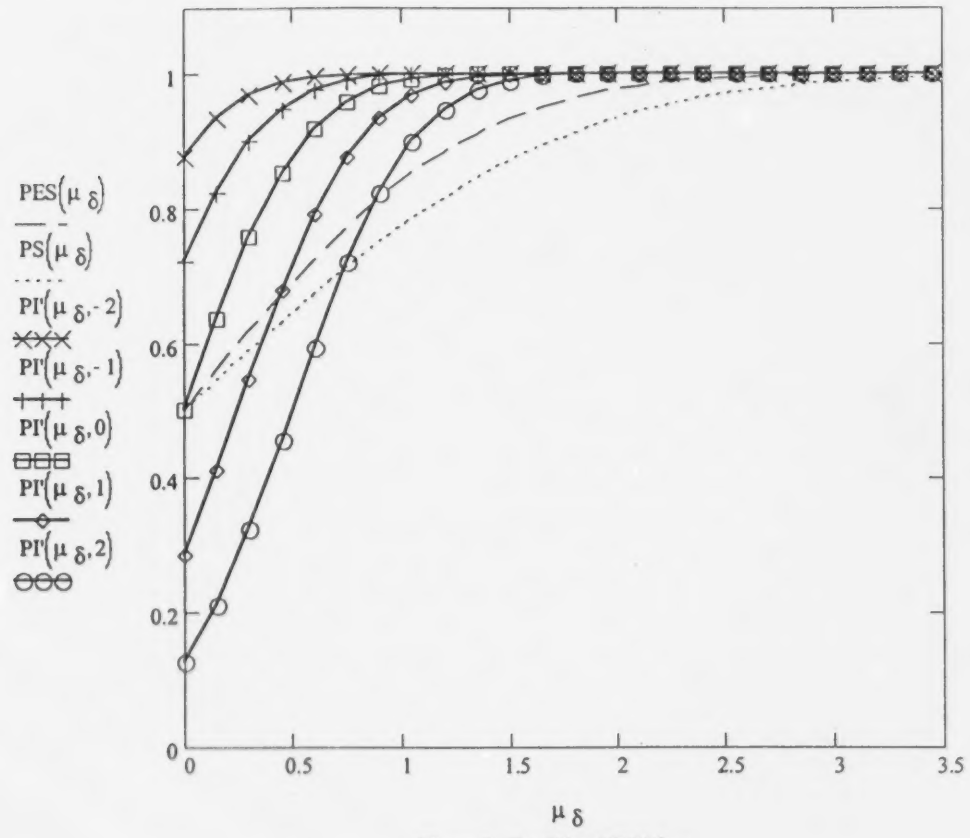

Figure 6. Conditional PI Values

For small values of μ_δ, as Figure 6 demonstrates, clients who are high on the criterion before a treatment effect is added have much less chance of improving than clients low on the criterion. For a client at $\varepsilon = 2$ with $\mu_\delta = 0.4$, the probability of improving is only 0.41, while for a client at $\varepsilon = -2$, the probability is 0.98. For example, if a client were two standard deviations above the mean on self-esteem and the treatment were not very effective (on the average), the chances are that he or she would not improve. From a counselor's perspective, this would be a very unusual client, but if you happened to be this client, you would want to know that an investment in counseling would be more likely than not to result in your self esteem decreasing.

The three simple models presented in this section serve only to define and compare the probabilities of different kinds of outcome. The models increase in complexity as treatment effects are first allowed to vary (Model II) and then allowed to relate to another variable (Model III). While they help in understanding, PES, PS, and PI, they do not begin, in our opinion, to represent real world complexity. In the next section, we begin to consider to the complexity we believe exists.

From simple to complex

When considering counseling outcome, counselors were told many years ago (Paul, 1967) to attend to the characteristics of the client and the counselor, the type of therapy, and the nature of the outcome sought. The number of variables that could be considered and the combinations of levels of those variables boggles the mind. Counseling researchers have proceeded by researching a few variables, some considered at only a few levels, and then analyzing the gathered data by postulating (at least implicitly) the most rudimentary model of counseling effect. The mean differences found in the research accumulate over time and eventually they are aggregated in a meta-analysis. The effect size found usually provides strong justification for the profession, which is important. But the simple probabilistic statement derived from the effect size provides virtually no basis for an individual client to decide to pursue counseling.

If we care about individual clients as consumers, we need to try to do better. How might we proceed? What can we do differently? First, we might consider forgetting about models. To predict how a client will do in counseling, it is not necessary to assume a model. A technique such as non-parametric regression can be used to do what we described at the beginning of this paper, namely to determine how a particular client will do in counseling by looking at the outcomes of similar clients. As background for the present paper, we carried out a number of investigations of non-parametric regression. It will not be a panacea, but it does offer a fresh approach worthy of further investigation. To move forward with non-parametric regression, or a similar technique, we will need to aggregate raw data rather than statistics from studies. This will cause problems, but the existence of the Internet would make pooling data easier than in the past. The database that would need to be created would require agreement on variables and a level of cooperation among researchers that would be challenging, but hopefully not impossible, to obtain.

As a last point, we must admit that the probability of improvement, PI, the probability we believe to be most relevant, is also the hardest to estimate given the way research data are often collected. Our next step is to consider research designs that would be particularly useful for collecting the data required to estimate values of PI.

References

Grissom, R. J. (1996). The magical number .7 ± .2: Meta-analysis of the probability of superior outcome in comparisons involving therapy, placebo, and control. Journal of Consulting and Clinical Psychology, 64, 973-982.

Paul, G. L. (1967). Strategy of outcome research in psychotherapy. Journal of Consulting Psychology, 31, 109-118.

Smith, M. L., & Glass, G. V. (1977). Meta-analysis of psychotherapy outcome studies. American Psychologist, 32, 752-760.

Smith, M. L., Glass, G. V., Miller, T, I. (1980). The benefits of psychothearpy. Baltimore: Johns Hopkins University Press.